Radiomics, Tumor Volume, and Blood Biomarkers for Early Prediction of Pseudoprogression in Patients with Metastatic Melanoma Treated with Immune Checkpoint Inhibition

Basler, Lucas ; Gabryś, Hubert S ; Hogan, Sabrina A ; Pavic, Matea ; Bogowicz, Marta ; Vuong, Diem ; Tanadini-Lang, Stephanie ; Förster, Robert ; Kudura, Ken ; Huellner, Martin W ; Dummer, Reinhard ; Guckenberger, Matthias ; Levesque, Mitchell P

Abstract: Purpose: We assessed the predictive potential of positron emission tomography (PET)/CT-based radiomics, lesion volume, and routine blood markers for early differentiation of pseudoprogression from true progression at 3 months. Experimental Design: 112 patients with metastatic melanoma treated with immune checkpoint inhibition were included in our study. Median follow-up duration was 22 months. 716 metastases were segmented individually on CT and 2[18F]fluoro-2-deoxy-D-glucose (FDG)-PET imaging at three timepoints: baseline (TP0), 3 months (TP1), and 6 months (TP2). Response was defined on a lesion-individual level (RECIST 1.1) and retrospectively correlated with FDG-PET/CT radiomic features and the blood markers LDH/S100. Seven multivariate prediction model classes were generated. Results: Two-year (median) overall survival, progression-free survival, and immune progression–free survival were 69% (not reached), 24% (6 months), and 42% (16 months), respectively. At 3 months, 106 (16%) lesions had progressed, of which 30 (5%) were identified as pseudoprogression at 6 months. Patients with pseudoprogressive lesions and without true progressive lesions had a similar outcome to responding patients and a significantly better 2-year overall survival of 100% (30 months), compared with 15% (10 months) in patients with true progressions/without pseudoprogression (P = 0.002). Patients with mixed progressive/pseudoprogressive lesions were in between at 53% (25 months). The blood prediction model (LDH+S100) achieved an AUC = 0.71. Higher LDH/S100 values indicated a low chance of pseudoprogression. Volume-based models: AUC = 0.72 (TP1) and AUC = 0.80 (delta-volume between TP0/TP1). Radiomics models (including/excluding volume-related features): AUC = 0.79/0.78. Combined blood-/volume model: AUC = 0.79. Combined blood/radiomics model (including volume-related features): AUC = 0.78. The combined blood/radiomics model (excluding volume-related features) performed best: AUC = 0.82. Conclusions: Noninvasive PET/CT-based radiomics, especially in combination with blood parameters, are promising biomarkers for early differentiation of pseudoprogression, potentially avoiding added toxicity or delayed treatment switch.

DOI: https://doi.org/10.1158/1078-0432.ccr-20-0020

Basler, Lucas; Gabryś, Hubert S; Hogan, Sabrina A; Pavic, Matea; Bogowicz, Marta; Vuong, Diem; Tanadini-Lang, Stephanie; Förster, Robert; Kudura, Ken; Huellner, Martin W; Dummer, Reinhard; Guckenberger, Matthias; Levesque, Mitchell P (2020). Radiomics, Tumor Volume, and Blood Biomarkers for Early Prediction of Pseudoprogression in Patients with Metastatic Melanoma Treated with Immune Checkpoint Inhibition. Clinical Cancer Research, 26(16):4414-4425.
DOI: https://doi.org/10.1158/1078-0432.ccr-20-0020

Title:

Radiomics, tumor volume and blood biomarkers for early prediction of pseudoprogression in metastatic melanoma patients treated with immune checkpoint inhibition

Authors and affiliations:

L. Basler[1], H.S. Gabryś[1], S. Hogan[2], M. Pavic[1], M. Bogowicz[1], D. Vuong[1], S. Tanadini-Lang[1], R. Förster[1], K. Kudura[3], M. Huellner[3], R. Dummer[2], M. Guckenberger[1*], M.P. Levesque[2*]

[1]Department of Radiation Oncology, University Hospital Zurich, University of Zurich, Zurich, Switzerland
[2]Department of Dermatology, University Hospital Zurich, University of Zurich, Zurich, Switzerland
[3]Department of Nuclear Medicine, University Hospital Zurich, University of Zurich, Zurich, Switzerland
* shared last authors

Running title:

Radiomics and biomarkers for prediction of pseudoprogression

Keywords:

1. Prediction of pseudoprogression
2. Metastatic melanoma patients
3. Immune checkpoint inhibition
4. Radiomics and delta-radiomics
5. Predictive biomarkers

Corresponding author: Mitch Levesque

University Hospital Zurich
Department of Dermatology
Gloriastrasse 31
CH 8091 Zürich
Switzerland

Phone: +41-(0) 43-253-3262
E-Mail: mitchell.levesque@usz.ch

Funding/Support:

CRC (Cancer Research Center) Funding program (CRC_13), Comprehensive Cancer Center Zurich, University Hospital Zurich, Zurich, Switzerland

Swiss National Fund (SNF 310030_170159)

European Training Network MELGEN funded consortium No. 641458

Conflicts of interest:

LB: none

HG: none

SH: none

MP: none

MB: none

DV: none

ST: none

RF: none

KK: none

MH: IIS grants and institutional grants from GE Healthcare, unrelated to the present study.

RD: Intermittent, project focused consulting and/or advisory relationships with Novartis, Merck Sharp & Dohme (MSD), Bristol-Myers Squibb (BMS), Roche, Amgen, Takeda, Pierre Fabre, Sun Pharma, Sanofi, Catalym, Second Genome outside the submitted work.

MG: none

MPL: Project-specific research support outside the scope of this paper from Roche, Novartis, and Bristol-Myers Squibb.

Translational relevance:

Immune checkpoint inhibitors (ICI) have revolutionized the treatment of metastatic melanoma patients. However, more than 50% of patients do not respond to ICI.

ICI response assessment is challenging, as novel response patterns, such as pseudoprogression (PP) are not considered in the response evaluation criteria in solid tumors (RECIST 1.1). An increase in tumor volume could be based on either true progressive disease (TPD) or on influx of immune-competent cells (PP). Early differentiation of PP and TPD is highly relevant in daily clinical decision-making, and predictive biomarkers are needed for better patient selection.

We could identify FDG-PET/CT-based radiomic and delta-radiomic features as novel imaging markers for early differentiation of PP from TPD. In addition, we could show that the routine blood markers LDH and S100 can contribute to PP prediction. A multi-modality approach of combined radiomics and blood marker-based prediction model at an early time-point of 3 months yielded the best performance. Thereby, added toxicity or delayed treatment switch in metastatic melanoma patients treated with ICI might be potentially avoided.

Abstract:

Purpose: We assessed the predictive potential of PET/CT-based radiomics, lesion volume, and routine blood markers for early differentiation of pseudoprogression from true progression at 3 months.

Experimental design: 112 metastatic melanoma patients treated with immune checkpoint inhibition. Median follow-up: 22 months. All 716 metastases were segmented individually on CT and FDG-PET imaging at 3 time-points: baseline (TP0), 3 months (TP1), 6 months (TP2). Response was defined on a lesion-individual level (RECIST 1.1) and retrospectively correlated with FDG-PET/CT radiomic-features and the blood-markers LDH/S100. Seven multivariate prediction model-classes were generated.

Results: 2-year (median) overall survival, progression-free survival and immune-progression-free survival were 69% (not reached), 24% (6 months) and 42% (16 months). At 3 months, 106 (16%) lesions had progressed, of which 30 (5%) were identified as pseudoprogression at 6 months. Patients with pseudoprogressive lesions and without true-progressive lesions had a similar outcome to responding patients and a significantly better 2-year overall survival of 100% (30 months), compared to 15% (10 months) in patients with true progressions/without pseudoprogression (p=0.002). Patients with mixed progressive/pseudoprogressive lesions were in between at 53% (25 months).

The blood prediction-model (LDH+S100) achieved an AUC=0.71. Higher LDH/S100 values indicated a low chance of pseudoprogression. Volume-based models: AUC=0.72 (TP1) and AUC=0.80 (delta-volume between TP0/TP1). Radiomics-models (including/excluding volume-related features): AUC=0.79/0.78. Combined blood/volume-model: AUC=0.79. Combined blood/radiomics-model (including volume-related features): AUC=0.78. The combined blood/radiomics-model (excluding volume-related features) performed best: AUC=0.82.

Conclusions: Non-invasive PET/CT-based radiomics, especially in combination with blood parameters, are promising biomarkers for early differentiation of pseudoprogression, potentially avoiding added toxicity or delayed treatment switch.

Background:

Immune checkpoint inhibitors (ICI) have revolutionized the treatment of metastatic melanoma patients and are guideline-recommended treatment standards (1–3). However, more than 50% of patients do not respond to ICI (4). Predictive biomarkers are needed for better patient selection. Lactate dehydrogenase (LDH) is one of the only established melanoma biomarkers with prognostic value for overall survival (OS) (5). S100 is associated with OS and response to ipilimumab; whereas, other blood-markers and clinical markers yielded mixed results (6–10). PD-L1 expression is insufficient for patient selection, as both PD-L1 positive and negative patients benefit from ICI (3). Tumor mutational burden (TMB), T-cell clonality, circulating tumor DNA (ctDNA), immune gene signatures, as well as T-cell-inflamed gene expression profile are promising, but currently unavailable in routine practice (11–19). Most of these factors do not have an application in response assessment and require at least minimally invasive diagnostics, rendering them challenging for repeated analysis.

ICI response assessment is especially challenging, as the response evaluation criteria in solid tumors (RECIST 1.1) do not consider novel response patterns, such as pseudoprogression (PP) (20,21). In the setting of ICI, a lesion enlargement could be caused by either true progressive disease (TPD) because of tumor growth or an influx of immune-competent cells, representing an effective anti-tumor immune response or pseudoprogression (22). Early differentiation of PP and TPD is highly relevant in daily clinical decision-making. The immune-related response criteria (irRC) (23) were the first to include these phenomena, followed by iRECIST (24). Both rely on confirmation follow-up imaging with potentially negative consequences. A treatment might be changed early because PP could be misinterpreted as TPD using only RECIST criteria or a switch to an effective treatment alternative could be delayed while waiting for confirmatory follow-up imaging. With newly proposed PECRIT (25) and PERCIMT (26) criteria, lesions increasing in size or the appearance of new lesions do not necessarily imply true progression, but may also be attributed to pseudoprogression. However, these criteria have not been validated in larger cohorts.

Non-invasive imaging is performed repetitively during the treatment course for continuous response assessment. Manual image assessment is, however, characterized by suboptimal accuracy as well as intra-observer and inter-observer variability (27,28). Consequently, there is a strong rationale for quantitative medical imaging analysis (i.e., radiomics). Recently, Sun et al. demonstrated a correlation between CT radiomic-signatures and the molecular CD8-cell expression in patients with different solid tumors treated with ICI, discriminating inflamed

tumors from non-inflamed tumors, which was associated with higher response rates at 3 and 6 months, as well as higher OS (29).

Our study aimed to identify novel imaging markers and blood markers for early differentiation of PP from TPD in metastatic melanoma patients treated with anti-PD-1 antibodies. The predictive potential of early single time-point radiomics and lesion volume as well as multi time-point delta-radiomics and delta-volume was assessed on a lesion-individual level using PET/CT imaging. In addition, the routine blood-markers LDH and S100 were assessed for PP prediction on a patient-individual level.

Material and Methods:

Patient Cohort

This is a single institution analysis of a deeply characterized cohort of 190 metastatic melanoma patients treated with either single checkpoint-inhibition (anti-PD-1) or dual checkpoint inhibition (anti-PD-1/anti-CTLA-4) between 2013 and 2019. Written informed consent was obtained from all patients and the study was approved by the local ethics committee (Kantonale Ethikkommission Zürich, approval number 2019-01012) in accordance with 'good clinical practice' (GCP) guidelines and the Declaration of Helsinki.

The following exclusion criteria were applied to allow for a standardized radiomics and outcome analysis: Lack of follow-up/baseline imaging; patients with only contrast-enhanced CT imaging (as most patients were staged/followed with non-enhanced PET/CT-imaging); patients with only brain metastases; patients presenting with only very small metastases at baseline (all baseline lesions <0.5 cc). The last exclusion criterion is based on multiple factors. Statistical comparisons between individual voxels of a defined lesion (e.g. radiomics analysis) can only be performed if an adequate number of voxels is present in the defined volume. This is limited by the resolution of the underlying imaging technology and the voxel size itself. In addition, all imaging-based analysis methods do require a minimum lesion size to perform measurements with sufficient accuracy and reliability.

Endpoints

On a lesion-individual level, response was defined using RECIST 1.1 criteria, comparing lesion diameter at three different time-points (TP): baseline (TP0), first follow-up at 3 months (TP1), second follow-up at 6 months (TP2). PP was defined as a diameter increase by ≥20% at TP1, followed by a decrease to <20% at TP2 compared to TP0. TPD was defined as an increase by ≥20% on both TP1 and TP2 compared to TP0.

The distribution of PP and TPD lesions was analyzed on a patient-level and all patients were classified into: (1) Patients with ≥1 PP lesion and no TPD lesions (PP-only); (2) Patients with ≥1 TPD and no PP lesions (TPD-only); (3) Patients presenting with both PP lesions and TPD lesions (mixed PP&TPD); (4) All other patients, who did not have a progressive lesion at TP1. For the lesion-level analysis and radiomics analysis, the appearance of new lesions was not taken into account in the patient stratification, as only lesions available at all three time-points were considered for these analyses. Overall survival (OS) was defined as the time from ICI treatment initiation to the date of death. Progression-free survival (PFS) and immune progression-free survival (iPFS) were defined as the time from ICI treatment initiation to the date of first progression/appearance of new lesions (PFS, RECIST(20)), or confirmed progression/appearance of new lesions (iPFS, iRECIST(24)).

Blood markers

LDH and S100 levels were measured at TP0 and TP1 and their relative change between TP1 and TP0 was calculated.

Imaging and lesion delineation

All imaging was performed at a single institution using standardized imaging-protocols. All subjects were injected with a body-weight-dependant and / or BMI-adapted FDG dose (2.0 - 3.5 MBq per kg). Scanning was performed on different scanners, partly with time-of-flight acquisition. PET image reconstructions used ordered subset expectation maximization together with point spread function modelling where available. The CT acquisition parameters were almost identical for all scanners and have been described previously in detail (30).

Based on the exclusion criteria, 112 patients with CT imaging for all three time-points were included. PET imaging for all time-points was available for 90 of these patients (80%). All lesions were manually segmented by two experienced clinicians based on a common protocol and consistent quality control at all time-points (**Figure 1**). A validation was performed for about 10% of all lesions, visually comparing the independently contoured lesions, demonstrating reproducibility across all lesion locations.

CT and PET images were coregistered rigidly and CT-based contours were propagated to the PET images. Any spatial geographic mismatch was manually corrected via shifting the CT-based contours to the corresponding lesion location in PET images. All lesions were classified into liver, lung, bone or soft tissue (including lymph nodes, cutaneous/subcutaneous and muscular lesions) metastases.

Extraction of radiomic features

The in-house developed radiomics software Z-Rad (31) written in Python programming language was used for preprocessing and extraction of radiomic features from medical images in agreement with the image biomarker standardization initiative (32). CT and PET images were resized to isotropic voxels of size 3.75 mm and 5 mm, respectively, corresponding to the lowest image resolution (image slice thickness) of the whole cohort. Three types of features, describing shape, intensity, and texture were extracted. No Hounsfield unit range limits were applied. As a preprocessing step before texture features extraction, image intensity values were discretized using a quantization step of 5 HU for CT and 0.25 SUV for PET. In total, 172 features per lesion were extracted for each imaging modality. The features were extracted for images taken at TP0 and TP1. The reproducibility of radiomic features was independently confirmed for a subset of lesions. Next, the relative (%) change in feature values between these two time-points was calculated which gave rise to delta-radiomic features.

Statistical analysis

To differentiate progressive lesions at TP1 into PP and TPD, a total of seven model classes were considered, including multi-modality approaches:

1) **Blood**: models based only on blood markers (LDH and S100).
2) **Volume**: models based only on the volume of individual metastases.
3) **Radiomics:** (including volume-related features) - models based on radiomic features.
4) **Radiomics:** (excluding volume-related features) - models based on radiomic features. Lesion volume and radiomic features correlated to lesion volume above Pearson's $r = 0.5$ were excluded.
5) **Blood and volume**: models based on blood markers and lesion volume.
6) **Blood and radiomics:** (excluding volume-related features) - models based on blood markers and on radiomic features. Lesion volume and radiomic features correlated to lesion volume above Pearson's $r = 0.5$ were excluded.
7) **Blood and radiomics:** (including volume-related features) - models based on blood markers and on radiomic features.

The model building procedure consisted of five steps. First, an unsupervised feature selection with the Pearson correlation coefficient was performed. In this procedure, pairwise correlation coefficients were calculated for all features and the correlated features were removed. Second, all features were scaled by subtracting the mean and dividing by the standard deviation. Third, a univariate supervised feature selection was done. For each

feature, an *F*-test was performed; only features significant after the false discovery rate correction with the Benjamini-Hochberg (33) procedure were kept. This step was followed by feature selection based on feature weights from a fitted model. For this reason, a logistic regression model regularized with elastic net was fit to the data. Features with the highest weights were selected. Finally, the classifier was fit to the data. This final model was based on logistic regression and regularized with L_2 penalty. The hyperparameters that were tuned are listed in **Table S1 (supplement)**.

Our models were trained, optimized, and tested in a setting of a nested cross-validation. The inner loop, used for model tuning, was a 3-times-repeated stratified 10-fold cross-validation. In total, 96 randomly generated hyperparameter samples were evaluated in model tuning with random search optimization (34). The outer loop, used for model testing, was a 3-times repeated stratified 5-fold cross-validation. The performance metric used was the area under the receiver operating characteristic curve (AUC).
The optimal cutoff was estimated for every model from the corresponding averaged ROC curve. The averaged ROC curves were calculated based on 15 ROC curves (3x5) from the outer loop of the nested cross-validation. The optimal cutoff was defined as the sensitivity and specificity at which the Youden's index (sensitivity + specificity - 1) was maximal (35). For the optimal cutoff, sensitivity, specificity, predictive values, and likelihood ratios with corresponding confidence intervals were estimated.

OS, PFS and iPFS were assessed and compared among PP-only, TPD-only, mixed PP&TPD groups, and all patients that did not progress at TP1. The landmark analysis method (36) has been used to avoid the guarantee-time bias. The landmark was set at the determination time-point of pseudoprogression and true progression (TP2). To avoid excluding patients who had their TP2 follow-up examination slightly earlier than 6 months after the onset of treatment, we chose 5 months for the landmark. Family-wise error rate (FWER) was controlled at the 0.05 level with the Holm-Bonferroni method for each type of survival (37).

Tumor burden, approximated by total tumor volume of metastatic lesions in a patient, was estimated at TP0, TP1 and TP2. New lesions that appeared at TP1 or TP2 were not considered in the estimation. Statistical significance of the differences between the groups was estimated with Mann-Whitney *U* test.

Software

For visualization, statistical analysis, model building, and model testing, the following open-source Python packages were used: HoloViews, Lifelines, Matplotlib, NumPy & SciPy, Pandas, Scikit-learn.

Results:

Patient characteristics:

Based on our exclusion criteria, 112 patients with a total of 716 metastases were included. A total of 2,061 metastases for both CT and PET imaging were individually segmented and analyzed. 645 (90%) of the 716 baseline lesions were either visible at all three time-points or had a complete remission at TP1 or TP2. The remaining 71 lesions had to be excluded due to either surgical removal or lack of follow-up imaging.

The analysis of the lesion locations, as well as the patient characteristics analysis was performed at baseline (TP0) and is therefore based on the 716 baseline metastases. The most frequent location was in soft tissue with 378 (52.8%) of all 716 baseline lesions. Other locations included 161 (22.5%) lung, 128 (17.9%) liver/spleen, 47 (6.6%) bone and 2 (0.3%) myocardial lesions. **Table 1** describes the patient characteristics for all groups.

The number of lesions per patient and lesion location were distributed equally between PP-only and TPD-only patients. Patients with mixed PP/TPD presented a higher mean number of 9.2 metastases per patient and had overall less soft tissue lesions compared to the other two groups, resembling a distribution closer to that of all patients combined. The percentage of patients receiving dual checkpoint-inhibition and prior anti-CTLA-4 treatment was evenly distributed among all groups. The overall percentage of patients having new lesions at TP1 and TP2 was 41% and 26%, respectively. PP-only patients presented with comparable numbers of 55% and 22%, respectively. TPD-only patients and mixed PP&TPD patients had a higher percentage of new lesions with 75%/73% (TP1) and 62%/64% (TP2), respectively.

Lesion-level analysis

In order to classify a single lesion into either pseudoprogression or true progression, a total of 3 time-points are necessary: baseline lesion diameter (TP0), diameter increase by ≥20% at 3 months (TP1), confirmation of either true progressive disease or pseudoprogression at 6 months (TP2). Therefore, the analysis of the individual lesion response per time-point was limited to the 645 lesions that were available at all 3 time-points.

Of these 645 lesions, 82 (13%) showed complete remission at the first follow-up at 3 months (TP1). Those with partial remission were 122 (19%), and 335 (52%) lesions were stable. Importantly, 106 (16%) lesions showed a progression by ≥20% at TP1, of which 30 lesions (4.7% of all lesions, 28.3% of progressive lesions) were defined as **pseudoprogression (PP)** at TP2. 76 lesions (11.8% of all lesions, 71.7% of progressive lesions) remained progressive throughout TP1 and TP2 and were classified as **true progression (TPD)**. The development of PP was not associated with metastasis location (p=0.40) or treatment type (single vs. dual ICI, p=0.12). **Figure 2** illustrates the changes in response between TP1 and TP2.

Pseudoprogression prediction models

A total of seven model classes for the prediction of pseudoprogression were considered, including multi-modality approaches. **Figure 3** shows the best performing ROC curves for all model modalities. The feature weights of the individual models, as well as partial dependence plots are provided in the supplement (**Figure S1 and S2**). **Table 2** provides an overview of all model metrics including AUC, sensitivity (true positive rate), specificity (true negative rate), positive/negative predictive value and positive/negative likelihood ratio.

Blood-based models

The first model was only based on the conventional blood markers LDH and S100 and achieved an AUC of 0.71 (sensitivity 0.69, specificity 0.67). Higher values of LDH and S100 indicated a low chance of pseudoprogression.

Volume-based models

As a second step, two models based solely on lesion volume (metastatic size) were generated. First, we created a model correlating pseudoprogression with the lesion volume at 3 months (TP1). This prediction model achieved an AUC of 0.72 (sensitivity 0.76, specificity 0.60). Smaller lesions were more likely to be pseudoprogressions. The second model was based on the relative (%) difference (delta-volume) between the lesion volume at 3 months (TP1) compared to baseline (TP0) and achieved and AUC of 0.80 (sensitivity 0.81, specificity 0.67). A large increase in lesion volume at TP1 was associated with a reduced likelihood of pseudoprogression.

Radiomics-based models

Four types of radiomics-based prediction models were generated and were based on either single time-point radiomics (TP1) or multi time-point delta-radiomics (relative difference between TP1 and TP0) with inclusion or exclusion of volume-related features. The best performing single time-point radiomics models achieved an AUC of 0.69 (CT, sensitivity

0.71, specificity 0.60) and 0.68 (PET, sensitivity 0.48, specificity 0.80), both performing worse compared to either the blood-based or volume-based models. The delta-radiomics models performed better and achieved an AUC of 0.79 (including volume-related features, sensitivity 0.81, specificity 0.67) based on the features mc-volume and fractal dimension and an AUC of 0.78 (excluding volume-related features, sensitivity 0.89, specificity 0.53) based on the feature CT coarseness. Both, an increase in fractal dimension and a decrease in coarseness are indicating that significant changes from a uniform/homogeneous to a nonuniform/heterogeneous texture are more likely to be true progressions.

Combined blood and volume-based model

The combined blood and volume-based prediction model, based on lesion-volume, LDH and S100 performed equally well compared to the volume-based model and better compared to the radiomics models, achieving an AUC of 0.79 (sensitivity 0.80, specificity 0.67).

Combined blood and radiomics-based model

A combined blood and radiomics-based prediction model (including volume-related features) achieved an AUC of 0.78 (sensitivity 0.84, specificity 0.67). The best performing model was, however, a combination of blood and radiomics-based prediction excluding volume-related features, which achieved an AUC of 0.82 (sensitivity 0.81, specificity 0.73). This prediction model was based on the LDH level at TP1 and the relative change of CT coarseness between TP1 and TP0. Larger values of LDH and a larger decrease in CT coarseness indicated a lower chance of pseudoprogression.

Distribution of PP and TPD lesions within patients

The 30 lesions with pseudoprogression were distributed across 20 patients with ≥1 PP lesion (median 1, range 1-5). The 76 lesions with TPD were distributed across a total of 27 patients. An overlap of 11 patients presented with mixed PP&TPD lesions (**Figure 4**).

Patient-level analysis

Median follow-up was 22 (14–32.5) months. For the whole cohort of 112 patients, median OS, PFS and iPFS were “not reached”, 6 and 16 months, respectively. Two-year OS, PFS and iPFS were 69%, 24% and 42%, respectively. **Table S2 (supplement)** summarizes the outcome of all groups.

In the landmark analysis, PP-only patients had a significantly longer median OS of 30 vs. 10 months (p=0.002, FWER=0.01) compared to TPD-only patients with a 2-year OS of 100% vs. 15%, and a better, however not statistically significant, iPFS of “not reached” vs. 7 months (p=0.014, FWER=0.058). Importantly, PP-only patients did not show a significant

difference in OS (p=0.934), PFS (p=0.500) and iPFS (p=0.557) compared to all other patients, who only had responding or stable lesions at TP1. Patients with mixed PP&TPD presented with a worse median OS of 25 vs. 30 months compared to PP-only patients (p=0.127), but had a longer median OS of 25 vs. 10 months compared to TPD-only (p=0.058, FWER=0.174). These differences were, however, not statistically significant under FWER=0.05.

As the difference between the TPD-only and mixed PP&TPD cohorts was quite striking, despite the fact that both presented with truly progressive lesions, we compared the overall tumor burden of both cohorts at all three time points. The mean baseline (TP0) tumor burden was higher in the mixed PP&TPD group (121 +/- 177 cc) compared to the TPD-only (86 +/- 90 cc) group. This reversed for TP1, where the TPD-only group presented with a higher mean tumor burden of 280 +/- 417 cc, compared to 133 +/- 175 cc in the mixed PP&TPD group, which was confirmed at TP2 with 383 +/- 542 cc (TPD-only) and 153 +/- 188 cc (mixed PP&TPD), respectively. These differences were, however, not statistically significant with *p*-values of p=0.639, p=0.786 and p=0.415 for TP0, TP1 and TP2, respectively.

Discussion:

Incidence of pseudoprogression

The incidence of pseudoprogression has been described to be 4-10% in most studies of metastatic melanoma patients treated with ICI (3,38–41). In our analysis, 8% of patients presented with pseudoprogressive lesions (PP-only). Our study is, however, the first to include a lesion-individual analysis of pseudoprogression, showing an overall rate of 4.7%. It is especially noteworthy that the 30 pseudoprogressive lesions in our analysis constitute almost one third (28.3%) of the 106 progressive lesions at TP1 and that the 9 PP-only patients constitute almost one third of patients with PD at TP1. This suggests that, although pseudoprogression might be a rare phenomenon overall, it appears to be quite common among patients with progressive disease.

The improved outcome of PP-only patients with an impressive two-year OS of 100% was comparable to patients who did not have any progressive lesion at TP1 and is in line with an improved OS of patients with pseudoprogression in other studies (15,39). TPD-only patients had a significantly shorter OS, iPFS and a drastically lower 2-year OS of only 14% indicating the need for an early change in treatment strategy. Patients with mixed PP&TPD showed a survival, which was worse compared to PP-only, but better compared to TPD-only patients,

although both groups had lesions with true progression and presented with comparable rates of new lesions at TP1 and TP2. This suggests that a lesion-individual response assessment might provide advantages for patient stratification. In addition, we assessed if the overall tumor burden could be responsible for the difference between both groups. Interestingly the baseline tumor burden was actually higher in the mixed PP&TPD group compared to the TPD-only group, suggesting that overall tumor burden may not be the only critical factor.

An explanation might be that mixed PP&TPD-patients could have had a slowly developing and/or resolving pseudoprogression of lesions that were still classified as PD at 6 months. Some studies described continued benefit from immunotherapy, despite confirmed progression according to iRECIST criteria (40–43). It could therefore be possible that some lesions that were defined as confirmed progression in our analysis may respond much later during treatment, as these delayed tumor response dynamics have been described for both melanoma and NSCLC patients by Nishino et al (43,44) and Hodi et al (39).

In addition, Kong et al. described potentially prolonged residual disease following anti-PD-1 treatment with metabolically inactive lesions on PET/CT imaging (45). This indicates that even iRECIST/irRC have limitations and that it might be necessary to further increase confirmation follow-up time, include additional imaging information (e.g. radiomics), analyze response more deeply on a lesion-individual level, and finally to search for additional biomarkers.

Radiomics

The potential of radiomics to predict treatment response and outcome has been recognized for many tumor types and imaging modalities (46,47). Tang et. al demonstrated that certain radiomic features (low CT intensity/high heterogeneity) predict infiltrating CD3+ lymphocytes and PD-L1 expression, associated with outcome in non-small-cell lung cancer (NSCLC) patients (48). As mentioned earlier, Sun et al. could predict ICI responders by baseline CT radiomic signatures (29). Trebeschi et al. also used baseline CT radiomics for response prediction of metastatic melanoma and NSCLC patients treated with ICI (49). While their model worked well in NSCLC patients, it performed poorly in melanoma patients, which the authors attributed to the large variety of first-line treatments prior to ICI. In our analysis, approximately half of the patients received first-line anti-PD-1-antibodies, while the other half was mainly pre-treated with anti-CTLA-4 antibodies, representing a much more homogeneous patient cohort. In addition, our analysis included a 50% larger number of 716 analyzed lesions (112 patients) compared to 483 lesions (80 patients) and an additional third time-point at 6 months.

A potential explanation why radiomic signatures are able to predict response or differentiate between PP and TPD are the previously described biological differences in lesions with either true progression or pseudoprogression (22). In the case of TPD, a lesion could consist of mainly tumor cells alone. For PP, the picture could be much more heterogeneous on a cellular level, with a plethora of different involved immune cells, such as dendritic cells, cytotoxic CD8+ T-cells, macrophages, natural killer cells and others. This could potentially also lead to differences in radiomic features on CT and PET imaging. Trebeschi et al. reported that lesions with a greater morphological heterogeneity were more likely to respond to immunotherapy (49). Our analysis shows that an increase in fractal dimension and a decrease in coarseness are more likely to be true progressions. Both indicate a change from a homogeneous to a heterogeneous texture. It might therefore be necessary to not only assess the heterogeneity of a lesion on a single time-point but also the relative change of heterogeneity between different time-points (delta-radiomics) to adequately assess the also dynamic biological processes of immune cell infiltration into the tumor. Together, this could provide a basis for the detection of differences in an in-depth radiomics analysis, although these results will need to be confirmed in future prospective clinical trials. Interestingly, the solely volume-based models showed a slightly higher performance compared to the solely radiomics-based models in our analysis. As response assessment is already based on tumor volume, volume-based prediction models may therefore provide a simple and practical approach.

Biomarkers

Surprisingly, biomarkers for ICI response assessment have not evolved substantially over the past decade. PD-L1 expression is one of the only clinically established biomarkers for predicting the response to ICI. However, some studies have also presented contradictory results, as both PD-L1 positive and negative patients were shown to benefit from ICI (50,51). Tumor-infiltrating lymphocytes (TILs), as well as neoantigens are other promising biomarkers and a higher number of TILs or neoantigens was associated with better outcomes in a variety of cancers (11,51–53). Serum biomarkers provide multiple advantages over pathology-based biomarkers, as they can be obtained non-invasively and repetitively throughout the treatment course (54). ctDNA was used to discriminate PP from TPD in two lung adenocarcinoma patients treated with ICI (55). Yoshimura et al. described a decrease in serum cytokeratin 19 fragment levels in the setting of pseudoprogression under ICI (56). Lee et al. performed an analysis in 125 metastatic melanoma patients treated with ICI, differentiating patients with PP from TPD via favorable ctDNA and unfavorable ctDNA profiles (15). 1-year OS was significantly higher in patients with favorable ctDNA (82%) compared to patients with unfavorable ctDNA (39%). Their numbers and percentages of

patients with TPD were almost identical to our cohort, and also a third (31%) of their patients with TPD at TP1 had PP. Tumor mutational burden (TMB) has also been assessed as a predictive biomarker, although two recent studies have shown that TMB was not associated with the efficacy of pembrolizumab in NSCLC patients (57,58). Other advanced biomarkers are being investigated including high-dimensional single-cell mass cytometry (CyTOF) which was shown to be able to predict the response to anti-PD-1 immunotherapy in metastatic melanoma patients (59). In our analysis, the solely blood-based model was the worst-performing of all model-classes with an AUC of 0.71 but blood in conjunction with either tumor-volume or radiomics, led to the best performing models, suggesting that multi-modality models may represent the best approach.

Multi-modality

It has been shown that combined multi-modality approaches for outcome prediction could perform better than single modality clinical, radiomics, or genomic data (60). We did not identify a study focusing on the non-invasive differentiation of pseudoprogression and true progression via radiomics and routine blood markers. The radiomics models of Sun et al. and Trebeschi et al. were solely based on CT imaging without PET imaging and neither included a multi time-point delta-radiomics analysis or a combined multi-modality model.

The predictive performance of FDG-PET/CT and blood markers for general response assessment in melanoma has only been compared in the setting of chemotherapy, surgery, or radiotherapy. Aukema et al. reported a modest 50% positive predictive value of S100 for recurrent disease and recommended FDG-PET/CT for confirmation of recurrences (61). Wieder et al. concluded that the diagnostic accuracy and prognostic power of PET/CT is superior to S100 (62). Strobel et al. reported that S100 alone was not suitable for response assessment in 37% of patients, since S100-values were normal prior and after treatment, whereas, PET/CT was suitable for all patients (63). Together, these results suggest that PET/CT imaging and the routine blood markers LDH and S100 could be combined for improved response assessment, but their use as a combined biomarker in immunotherapy has not been examined systematically.

We could show that the combination of delta-radiomics and LDH was the best performing model, performing better than all other model-classes. The multi-modality model indicated that high levels of LDH and lesions with a large decrease in CT coarseness have a low chance of being a pseudoprogression.

Limitations

There are some limitations of our study, including the retrospective character of the analysis and the overall limited number of pseudoprogressive lesions and patients in the individual groups (PP-only, TPD-only, PP&TPD).

No external validation of our model has been performed. The most straightforward way to select and test the optimal model is to split the data set into three parts: the training set, the validation set, and the test set. The training set is used to fit the models to the data, the validation set is used to select the optimal model, and the test set is used to estimate the expected predictive performance in an independent data set, that is, generalization performance. This is a recommended approach in large data sets. However, in small data sets the way the data set is split can significantly affect the performance scores in both validation and testing contributing to a high variance of the performance estimate. A common approach to reduce this variance is cross-validation. In cross-validation, the process of splitting the data to training and validation parts is repeated multiple times, reducing variance by averaging the performance scores (64). However, it has been shown by Stone (65) that cross-validation can lead to optimistically biased performance estimates when it is used for both model selection and predictive performance assessment (64). A solution to this problem was proposed by Varma and Simon (66) who showed that nesting two cross-validation loops, where the inner loop selects the optimal model and the outer loop performs the model assessment, mitigates this issue resulting in an estimation of the true generalization performance with very low bias. In small data sets, nested cross-validation is therefore superior to a training/validation/test split or a single cross-validation as it provides an almost unbiased estimate of true generalization performance with low variance (67).

However, small sample size may have prevented the multivariate models from reaching full capacity. A larger patient cohort could therefore lead to an improved performance of the multivariate models. Another limitation is the heterogenous PET/CT acquisition and reconstruction techniques, which is owing to the retrospective nature of our study. Nevertheless, despite these limitations, our study is by far the largest lesion-level analysis of metastatic melanoma patients treated with ICI and the only one to include 3 separate time-points, blood biomarkers, delta-radiomics, and PET/CT imaging.

Conclusion:

This study reports the first multi-modality radiomics analysis using FDG-PET/CT imaging in the setting of immune checkpoint-inhibition. It is also the first to include a multiple time-point delta radiomics analysis in a multi-modality prediction model in conjunction with the blood

markers LDH and S100. Non-invasive PET/CT-based radiomics and LDH/S100 are promising biomarkers for early differentiation of pseudoprogression from true progression at an early time-point of 3 months, which might help reduce added toxicity or delayed treatment switch in metastatic melanoma patients treated with immune checkpoint inhibitors.

Acknowledgements:

Author Contributions:

Prof. Levesque and Prof. Guckenberger have full access to all data used in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Levesque, Dummer, Guckenberger, Basler, Foerster, Bogowicz, Tanadini.

Acquisition, analysis, or interpretation of data: Basler, Gabryś.

Drafting of the manuscript: All authors.

Critical revision of the manuscript for important intellectual content: All authors.

Statistical analysis: Gabryś.

Obtained funding: Basler, Gabryś, Hogan, Levesque, Guckenberger

Supervision: Levesque, Guckenberger.

References

1. Hodi FS, O'Day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, et al. Improved Survival with Ipilimumab in Patients with Metastatic Melanoma. N Engl J Med. Massachusetts Medical Society; 2010;363:711–23.
2. Larkin J, Chiarion-Sileni V, Gonzalez R, Grob JJ, Cowey CL, Lao CD, et al. Combined Nivolumab and Ipilimumab or Monotherapy in Untreated Melanoma. N Engl J Med. Massachusetts Medical Society; 2015;373:23–34.
3. Robert C, Schachter J, Long GV, Arance A, Grob JJ, Mortier L, et al. Pembrolizumab versus Ipilimumab in Advanced Melanoma. N Engl J Med. Massachusetts Medical Society; 2015;372:2521–32.
4. Ascierto PA, Marincola FM. 2015: The Year of Anti-PD-1/PD-L1s Against Melanoma and Beyond. EBioMedicine. 2015;2:92–3.
5. Gershenwald JE, Scolyer RA, Hess KR, Sondak VK, Long GV, Ross MI, et al. Melanoma staging: Evidence-based changes in the American Joint Committee on Cancer eighth edition cancer staging manual. CA Cancer J Clin. 2017;67:472–92.
6. Kelderman S, Heemskerk B, van Tinteren H, van den Brom RRH, Hospers GAP, van den Eertwegh AJM, et al. Lactate dehydrogenase as a selection criterion for ipilimumab treatment in metastatic melanoma. Cancer Immunol Immunother. 2014;63:449–58.

7. Kaskel P, Berking C, Sander S, Volkenandt M, Peter RU, Krähn G. S-100 protein in peripheral blood: a marker for melanoma metastases: a prospective 2-center study of 570 patients with melanoma. J Am Acad Dermatol. 1999;41:962–9.

8. Simeone E, Gentilcore G, Giannarelli D, Grimaldi AM, Caracò C, Curvietto M, et al. Immunological and biological changes during ipilimumab treatment and their potential correlation with clinical response and survival in patients with advanced melanoma. Cancer Immunol Immunother. 2014;63:675–83.

9. Hopkins AM, Rowland A, Kichenadasse G, Wiese MD, Gurney H, McKinnon RA, et al. Predicting response and toxicity to immune checkpoint inhibitors using routinely available blood and clinical markers. Br J Cancer. 2017;117:913–20.

10. Hogan SA, Levesque MP, Cheng PF. Melanoma Immunotherapy: Next-Generation Biomarkers. Front Oncol. frontiersin.org; 2018;8:178.

11. Galon J, Mlecnik B, Bindea G, Angell HK, Berger A, Lagorce C, et al. Towards the introduction of the “Immunoscore”in the classification of malignant tumours. J Pathol. Wiley Online Library; 2014;232:199–209.

12. Gibney GT, Weiner LM, Atkins MB. Predictive biomarkers for checkpoint inhibitor-based immunotherapy. Lancet Oncol. Elsevier; 2016;17:e542–51.

13. Cristescu R, Mogg R, Ayers M, Albright A, Murphy E, Yearley J, et al. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. Science [Internet]. 2018;362. Available from: http://dx.doi.org/10.1126/science.aar3593

14. Ott PA, Bang Y-J, Piha-Paul SA, Razak ARA, Bennouna J, Soria J-C, et al. T-Cell-Inflamed Gene-Expression Profile, Programmed Death Ligand 1 Expression, and Tumor Mutational Burden Predict Efficacy in Patients Treated With Pembrolizumab Across 20 Cancers: KEYNOTE-028. J Clin Oncol. 2019;37:318–27.

15. Lee JH, Long GV, Menzies AM, Lo S, Guminski A, Whitbourne K, et al. Association Between Circulating Tumor DNA and Pseudoprogression in Patients With Metastatic Melanoma Treated With Anti-Programmed Cell Death 1 Antibodies. JAMA Oncol. 2018;4:717–21.

16. Lee JH, Long GV, Boyd S, Lo S, Menzies AM, Tembe V, et al. Circulating tumour DNA predicts response to anti-PD1 antibodies in metastatic melanoma. Ann Oncol. 2017;28:1130–6.

17. Hwang S, Kwon A-Y, Jeong J-Y, Kim S, Kang H, Park J, et al. Immune gene signatures for predicting durable clinical benefit of anti-PD-1 immunotherapy in patients with non-small cell lung cancer. Sci Rep. 2020;10:643.

18. Jamieson NB, Maker AV. Gene-expression profiling to predict responsiveness to immunotherapy. Cancer Gene Ther. 2017;24:134–40.

19. Jiang P, Gu S, Pan D, Fu J, Sahu A, Hu X, et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. Nat Med. 2018;24:1550–8.

20. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45:228–47.

21. Borcoman E, Kanjanapan Y, Champiat S, Kato S, Servois V, Kurzrock R, et al. Novel patterns of response under immunotherapy. Ann Oncol. 2019;30:385–96.

22. Beer L, Hochmair M, Prosch H. Pitfalls in the radiological response assessment of immunotherapy. Memo. 2018;11:138–43.

23. Wolchok JD, Hoos A, O’Day S, Weber JS, Hamid O, Lebbé C, et al. Guidelines for the evaluation of immune therapy activity in solid tumors: immune-related response criteria. Clin Cancer Res. 2009;15:7412–20.

24. Seymour L, Bogaerts J, Perrone A, Ford R, Schwartz LH, Mandrekar S, et al. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. Lancet Oncol. Elsevier; 2017;18:e143–52.

25. Cho SY, Lipson EJ, Im H-J, Rowe SP, Gonzalez EM, Blackford A, et al. Prediction of Response to Immune Checkpoint Inhibitor Therapy Using Early-Time-Point 18F-FDG PET/CT Imaging in Patients with Advanced Melanoma. J Nucl Med. 2017;58:1421–8.

26. Anwar H, Sachpekidis C, Winkler J, Kopp-Schneider A, Haberkorn U, Hassel JC, et al. Absolute number of new lesions on 18F-FDG PET/CT is more predictive of clinical response than SUV changes in metastatic melanoma patients receiving ipilimumab. Eur J Nucl Med Mol Imaging. 2018;45:376–83.

27. Erasmus JJ, Gladish GW, Broemeling L, Sabloff BS, Truong MT, Herbst RS, et al. Interobserver and intraobserver variability in measurement of non--small-cell carcinoma lung lesions: implications for assessment of tumor response. J Clin Oncol. American Society of Clinical Oncology; 2003;21:2574–82.

28. Hopper KD, Kasales CJ, Van Slyke MA, Schwartz TA, TenHave TR, Jozefiak JA. Analysis of interobserver and intraobserver variability in CT tumor measurements. AJR Am J Roentgenol. 1996;167:851–4.

29. Sun R, Limkin EJ, Vakalopoulou M, Dercle L, Champiat S, Han SR, et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. Lancet Oncol. 2018;19:1180–91.

30. Huellner MW, Appenzeller P, Kuhn FP, Husmann L, Pietsch CM, Burger IA, et al. Whole-body nonenhanced PET/MR versus PET/CT in the staging and restaging of cancers: preliminary observations. Radiology. 2014;273:859–69.

31. Bogowicz M, Leijenaar RTH, Tanadini-Lang S. Post-radiochemotherapy PET radiomics in head and neck cancer–The influence of radiomics implementation on the reproducibility of local control tumor …. Radiother Oncol [Internet]. Elsevier; 2017; Available from: https://www.sciencedirect.com/science/article/pii/S0167814017326634

32. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative [Internet]. arXiv [cs.CV]. 2016. Available from: http://arxiv.org/abs/1612.07003

33. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. Wiley Online Library; 1995;57:289–300.

34. Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. J Mach Learn Res. 2012;13:281–305.

35. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3:32–5.

36. Dafni U. Landmark analysis at the 25-year landmark point. Circ Cardiovasc Qual Outcomes. 2011;4:363–71.

37. Holm S. A simple sequentially rejective multiple test procedure. Scand Stat Theory Appl [Internet]. JSTOR; 1979; Available from: https://www.jstor.org/stable/4615733

38. Chiou VL, Burotto M. Pseudoprogression and Immune-Related Response in Solid Tumors. J Clin Oncol. 2015;33:3541–3.

39. Hodi FS, Hwu W-J, Kefford R, Weber JS, Daud A, Hamid O, et al. Evaluation of Immune-Related Response Criteria and RECIST v1.1 in Patients With Advanced Melanoma Treated With Pembrolizumab. J Clin Oncol. 2016;34:1510–7.

40. Long GV, Weber JS, Larkin J, Atkinson V, Grob J-J, Schadendorf D, et al. Nivolumab for Patients With Advanced Melanoma Treated Beyond Progression: Analysis of 2 Phase 3 Clinical Trials. JAMA Oncol. 2017;3:1511–9.

41. Weber JS, D'Angelo SP, Minor D, Hodi FS, Gutzmer R, Neyns B, et al. Nivolumab versus chemotherapy in patients with advanced melanoma who progressed after anti-CTLA-4 treatment (CheckMate 037): a randomised, controlled, open-label, phase 3 trial. Lancet Oncol. 2015;16:375–84.

42. Song P, Zhang J, Shang C, Zhang L. Curative effect assessment of immunotherapy for non-small cell lung cancer: The "blind area" of Immune Response Evaluation Criteria in Solid Tumors (iRECIST). Thorac Cancer. 2019;10:587–92.

43. Nishino M, Giobbie-Hurder A, Manos MP, Bailey N, Buchbinder EI, Ott PA, et al. Immune-Related Tumor Response Dynamics in Melanoma Patients Treated with Pembrolizumab: Identifying Markers for Clinical Outcome and Treatment Decisions. Clin Cancer Res. 2017;23:4671–9.

44. Nishino M, Dahlberg SE, Adeni AE, Lydon CA, Hatabu H, Jänne PA, et al. Tumor Response Dynamics of Advanced Non-small Cell Lung Cancer Patients Treated with PD-1 Inhibitors: Imaging Markers for Treatment Outcome. Clin Cancer Res. 2017;23:5737–44.

45. Kong BY, Menzies AM, Saunders CAB, Liniker E, Ramanujam S, Guminski A, et al. Residual FDG-PET metabolic activity in metastatic melanoma patients with prolonged response to anti-PD-1 therapy. Pigment Cell Melanoma Res. 2016;29:572–7.

46. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14:749–62.

47. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.

48. Tang C, Hobbs B, Amer A, Li X, Behrens C, Canales JR, et al. Development of an Immune-Pathology Informed Radiomics Model for Non-Small Cell Lung Cancer. Sci Rep. 2018;8:1922.

49. Trebeschi S, Drago SG, Birkbak NJ, Kurilova I, Călin AM, Pizzi AD, et al. Predicting Response to Cancer Immunotherapy using Non-invasive Radiomic Biomarkers. Ann Oncol [Internet]. 2019; Available from: http://dx.doi.org/10.1093/annonc/mdz108

50. Daud AI, Wolchok JD, Robert C, Hwu W-J, Weber JS, Ribas A, et al. Programmed Death-Ligand 1 Expression and Response to the Anti-Programmed Death 1 Antibody Pembrolizumab in Melanoma. J Clin Oncol. 2016;34:4102–9.

51. Nishino M, Ramaiya NH, Hatabu H, Hodi FS. Monitoring immune-checkpoint blockade: response evaluation and biomarker development. Nat Rev Clin Oncol. 2017;14:655–68.

52. Tumeh PC, Harview CL, Yearley JH, Shintaku IP, Taylor EJM, Robert L, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. Nature. 2014;515:568–71.

53. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. Science. 2015;348:69–74.

54. Weide B, Martens A, Hassel JC, Berking C, Postow MA, Bisschop K, et al. Baseline Biomarkers for Outcome of Melanoma Patients Treated with Pembrolizumab. Clin Cancer Res. 2016;22:5487–96.

55. Guibert N, Mazieres J, Delaunay M, Casanova A, Farella M, Keller L, et al. Monitoring of KRAS-mutated ctDNA to discriminate pseudo-progression from true progression during anti-PD-1 treatment of lung adenocarcinoma. Oncotarget. 2017;8:38056–60.

56. Yoshimura A, Takumi C, Tsuji T, Hamashima R, Shiotsu S, Yuba T, et al. Pulmonary pleomorphic carcinoma with pseudoprogression during nivolumab therapy and the usefulness of tumor markers: A case report. Clin Case Rep. 2018;6:1338–41.

57. Langer C, Gadgeel S, Borghaei H, Patnaik A, Powell S, Gentzler R, et al. OA04.05 KEYNOTE-021: TMB and Outcomes for Carboplatin and Pemetrexed With or Without Pembrolizumab for Nonsquamous NSCLC. J Thorac Oncol. Elsevier; 2019;14:S216.

58. Garassino M, Rodriguez-Abreu D, Gadgeel S, Esteban E, Felip E, Speranza G, et al. OA04.06 Evaluation of TMB in KEYNOTE-189: Pembrolizumab Plus Chemotherapy vs Placebo Plus Chemotherapy for Nonsquamous NSCLC. J Thorac Oncol. Elsevier; 2019;14:S216–7.

59. Krieg C, Nowicka M, Guglietta S, Schindler S, Hartmann FJ, Weber LM, et al. High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. Nat. Med. 2018. page 144–53.

60. Grossmann P, Stringfield O, El-Hachem N, Bui MM, Rios Velazquez E, Parmar C, et al. Defining the biological basis of radiomic phenotypes in lung cancer. Elife [Internet]. 2017;6. Available from: http://dx.doi.org/10.7554/eLife.23421

61. Aukema TS, Olmos RAV, Korse CM, Kroon BBR, Wouters MWJM, Vogel WV, et al. Utility of FDG PET/CT and brain MRI in melanoma patients with increased serum S-100B level during follow-up. Ann Surg Oncol. 2010;17:1657–61.

62. Wieder HA, Tekin G, Rosenbaum-Krumme S, Klode J, Altenbernd J, Bockisch A, et al. 18FDG-PET to assess recurrence and long term survival in patients with malignant melanoma. Nuklearmedizin. 2013;52:198–203.

63. Strobel K, Skalsky J, Steinert HC, Dummer R, Hany TF, Bhure U, et al. S-100B and FDG-PET/CT in therapy response assessment of melanoma patients. Dermatology. 2007;215:192–201.

64. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer Science & Business Media; 2009.

65. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. J R Stat Soc Series B Stat Methodol. 1974;36:111–33.

66. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91.

67. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res. 2010;11:29.

Figure Legends

Fig. 1 Contouring and response assessment of each individual lesion On a lesion-individual level, response was defined using RECIST 1.1 criteria, comparing lesion diameter at three different time-points (TP): baseline (TP0), at the first follow-up at 3 months (TP1), as well as the second follow-up at 6 months (TP2). PP was defined as diameter increase by ≥20% at TP1, followed by a decrease to <20% at TP2 compared to TP0. TPD was defined as a consistent increase by ≥20% at TP1 and TP2. All metastatic lesions were manually segmented at all time-points.

Fig. 2 Change of individual lesion response between all time-points Individual change of response between baseline (TP0), 3 months (TP1) and 6 months (TP2) for all lesions that were available at all 3 time-points (n = 645). 106 (16%) progressive lesions (PD) were identified at TP1, of which 30 changed to either complete response (CR), partial response (PR) or stable disease (SD) at TP2, representing pseudoprogression (PP). 76

(71.7%) lesions remained progressive throughout TP1 and TP2 and were classified as true progressive disease (TPD).

Fig. 3 Multivariate models for prediction of pseudoprogression. AUC curves for the best performing models of all seven model-classes. **A**. blood-based model. **B**. volume-based model. **C**. radiomics-based model (incl. volume-related features). **D**. radiomics-based model (excl. volume-related features). **E**. combined blood & volume-based model. **F**. combined blood & radiomics-based model (incl. volume-related features). **G**. combined blood & radiomics-based model (excl. volume-related features), which was the best performing model and achieved an AUC of 0.82 (sensitivity = 0.81 ± 0.13, specificity = 0.73 ± 0.35). This prediction model is based on the LDH level at TP1 and the relative change of CT coarseness between TP1 and TP0. Larger values of LDH and a larger decrease in CT coarseness indicated a lower chance of pseudoprogression.

Fig. 4 Distribution of pseudoprogression and true progression and association with OS, PFS and iPFS A. 106 progressive lesions were present at TP1, 30 lesions with pseudoprogression were distributed across 20 patients with ≥1 PP-lesion. The 76 lesions with true progression were distributed across a total of 27 patients. An overlap of 11 patients presented with mixed PP&TPD lesions. Overall survival (B), progression-free survival (C) and immune-progression-free survival (D) of the different groups. The landmark was set at 5 months. PP-only patients had a significantly longer median OS of 30 vs. 10 months (p=0.002, FWER=0.010) compared to TPD-only patients with a 2-year OS of 100% vs. 15%.

Patient Characteristics	All patients	Patients with progressive lesions at TP1		
		Pseudo-progression only (PP-only)	True progression only (TPD-only)	Mixed pseudo-progression and true progression (PP&TPD)
General				
Total (N, %)	112 (100%)	9 (8%)	16 (14%)	11 (10%)
Female (N, %)	34 (30.4%)	3 (33.3%)	8 (50%)	6 (54.5%)
Male (N, %)	78 (69.6%)	6 (66.7%)	8 (50%)	5 (45.5 %)
Median age (years, IQR)	69 (55–76)	74 (64–78)	57.5 (50.5–69)	61 (51.5–73)
Treatment information				
Single Checkpoint Inhibition (N, %)	95 (84.8%)	7 (77.8%)	14 (87.5%)	10 (90.9%)
Dual Checkpoint Inhibition (N, %)	17 (15.2%)	2 (22.2%)	2 (12.5%)	1 (9.1%)
Prior treatments (cumulative):				
Total	69 (61%)	5 (55%)	10 (62%)	8 (72%)
Ipilimumab	53 (76.8%)	4 (80%)	7 (70%)	6 (75%)
Ipilimumab + Nivolumab	3 (4.4%)	0 (0%)	0 (0%)	0 (0%)
BRAF-Inhibitor	2 (2.9%)	0 (0%)	1 (10%)	0 (0%)
MEK-Inhibitor	4 (5.8%)	0 (0%)	0 (0%)	1 (12.5%)
Chemotherapy	7 (10.1%)	1 (20%)	2 (20%)	1 (12.5%)
Lesion details				
Total number of lesions (baseline)	716 (100%)	40 (5.6%)	93 (13.0%)	101 (14.1%)
Mean number of lesions per patient	6.4	4.5	5.8	9.2
1 = soft tissue	378 (52.8%)	29 (72.5%)	66 (71.0%)	50 (49.5%)
2 = lung	161 (22.5%)	6 (15.0%)	12 (12.9%)	23 (22.8%)
3 = liver/spleen	128 (17.9%)	5 (12.5%)	9 (9.7%)	15 (14.9%)
4 = bone	47 (6.6%)	0 (0%)	6 (6.5%)	13 (12.9%)
5 = heart	2 (0.3%)	0 (0%)	0 (0%)	0 (0%)
Patients with new lesions at TP1	46 (41%)	5 (55%)	12 (75%)	8 (73%)
Patients with new lesions at TP2	29 (26%)	2 (22%)	10 (62%)	7 (64%)

Table 1. Patient characteristics Patient characteristics of all included patients (n=112), as well as the defined groups: PP-only, TPD-only and mixed PP&TPD.

Model classes	Included parameters / features	Area under curve (AUC) (SD)	Sensitivity (TPR) (95% CI)	Specificity (TNR) (95% CI)	Positive predictive value (PPV) (95% CI)	Negative predictive value (NPV) (95% CI)	Positive likelihood ratio (LR+) (95% CI)	Negative likelihood ratio (LR-) (95% CI)
Blood	LDH, S-100B	0.71 (± 0.07)	0.69 (0.52, 0.85)	0.67 (0.56, 0.77)	0.45 (0.30, 0.59)	0.84 (0.75, 0.93)	2.06 (1.38, 3.07)	0.47 (0.27, 0.82)
Volume (TP1)	CT volume	0.72 (± 0.13)	0.76 (0.60, 0.91)	0.60 (0.49, 0.71)	0.42 (0.29, 0.56)	0.86 (0.77, 0.95)	1.87 (1.33, 2.63)	0.41 (0.21, 0.79)
Delta-volume (TP1 vs. TP0)	ΔCT volume	0.80 (± 0.10)	0.81 (0.67, 0.95)	0.67 (0.56, 0.77)	0.49 (0.35, 0.63)	0.90 (0.82, 0.98)	2.43 (1.69, 3.49)	0.28 (0.13, 0.60)
Radiomics (TP1 - CT)	CT center mass shift	0.69 (± 0.12)	0.71 (0.55, 0.87)	0.60 (0.49, 0.71)	0.41 (0.28, 0.54)	0.84 (0.74, 0.94)	1.76 (1.23, 2.51)	0.48 (0.27, 0.88)
Radiomics (TP1 - PET)	PET information measures of correlation 2, PET large zone high grey level emphasis	0.68 (± 0.13)	0.48 (0.30, 0.66)	0.80 (0.71, 0.89)	0.48 (0.30, 0.66)	0.79 (0.70, 0.89)	2.37 (1.32, 4.24)	0.65 (0.46, 0.94)
Delta-radiomics (TP1 vs. TP0) (including volume-related features)	ΔCT volume, ΔCT fractal-dimension	0.79 (± 0.09)	0.81 (0.67, 0.95)	0.67 (0.56, 0.77)	0.49 (0.35, 0.63)	0.90 (0.82, 0.98)	2.43 (1.69, 3.49)	0.28 (0.13, 0.60)
Delta-radiomics (TP1 vs. TP0) (excluding volume-related features)	ΔCT coarseness	0.78 (± 0.08)	0.89 (0.78, 1.00)	0.53 (0.41, 0.64)	0.42 (0.30, 0.55)	0.92 (0.84, 1.00)	1.87 (1.43, 2.45)	0.21 (0.08, 0.60)
Blood & Volume	LDH, S-100B, ΔCT volume	0.79 (± 0.08)	0.80 (0.66, 0.94)	0.67 (0.56, 0.77)	0.49 (0.35, 0.63)	0.89 (0.81, 0.97)	2.40 (1.67, 3.46)	0.30 (0.14, 0.62)
Blood & Radiomics (including volume-related features)	LDH, ΔCT volume, ΔCT fractal-dimension, ΔCT 10th percentile	0.78 (± 0.09)	0.84 (0.71, 0.97)	0.67 (0.56, 0.77)	0.50 (0.36, 0.64)	0.92 (0.84, 0.99)	2.53 (1.78, 3.61)	0.23 (0.10, 0.55)
Blood & Radiomics (excluding volume-related features)	LDH, ΔCT coarseness	0.82 (± 0.09)	0.81 (0.67, 0.95)	0.73 (0.63, 0.83)	0.54 (0.39, 0.69)	0.91 (0.83, 0.98)	2.97 (1.98, 4.46)	0.26 (0.12, 0.55)

Table 2. Model metrics for all seven model-classes. Overview of all model metrics including AUC (± standard deviation), sensitivity (true positive rate, 95% confidence intervals), specificity (true negative rate, 95% confidence intervals), positive/negative predictive value (PPV/NPV, 95% confidence intervals) and positive/negative likelihood ratio (LR+/LR-, 95% confidence intervals).

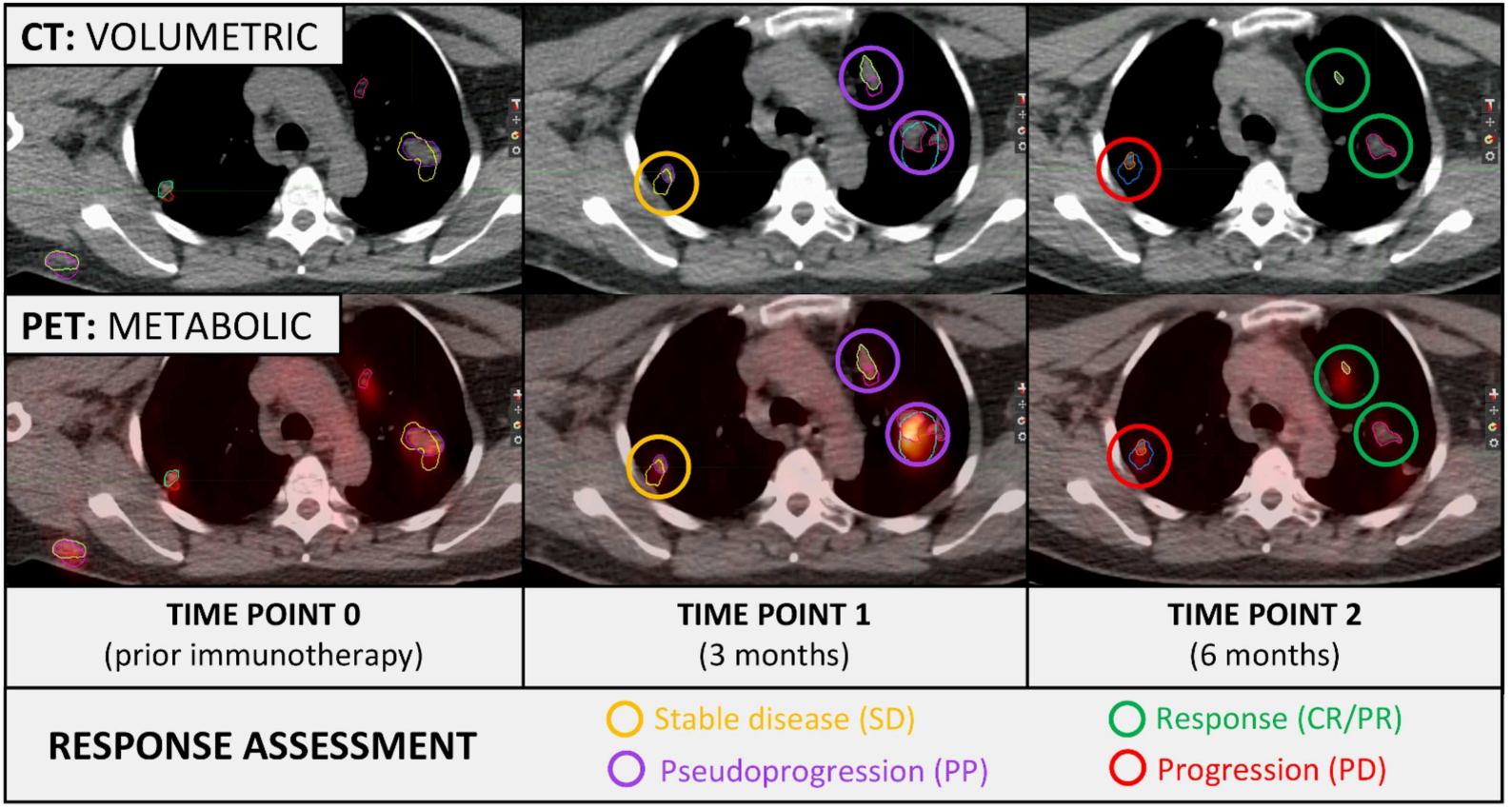

Fig. 1 Contouring and response assessment of each individual lesion On a lesion-individual level, response was defined using RECIST 1.1 criteria, comparing lesion diameter at three different time-points (TP): baseline (TP0), at the first follow-up at 3 months (TP1), as well as the second follow-up at 6 months (TP2). PP was defined as diameter increase by ≥20% at TP1, followed by a decrease to <20% at TP2 compared to TP0. TPD was defined as a consistent increase by ≥20% at TP1 and TP2. All metastatic lesions were manually segmented at all time-points.

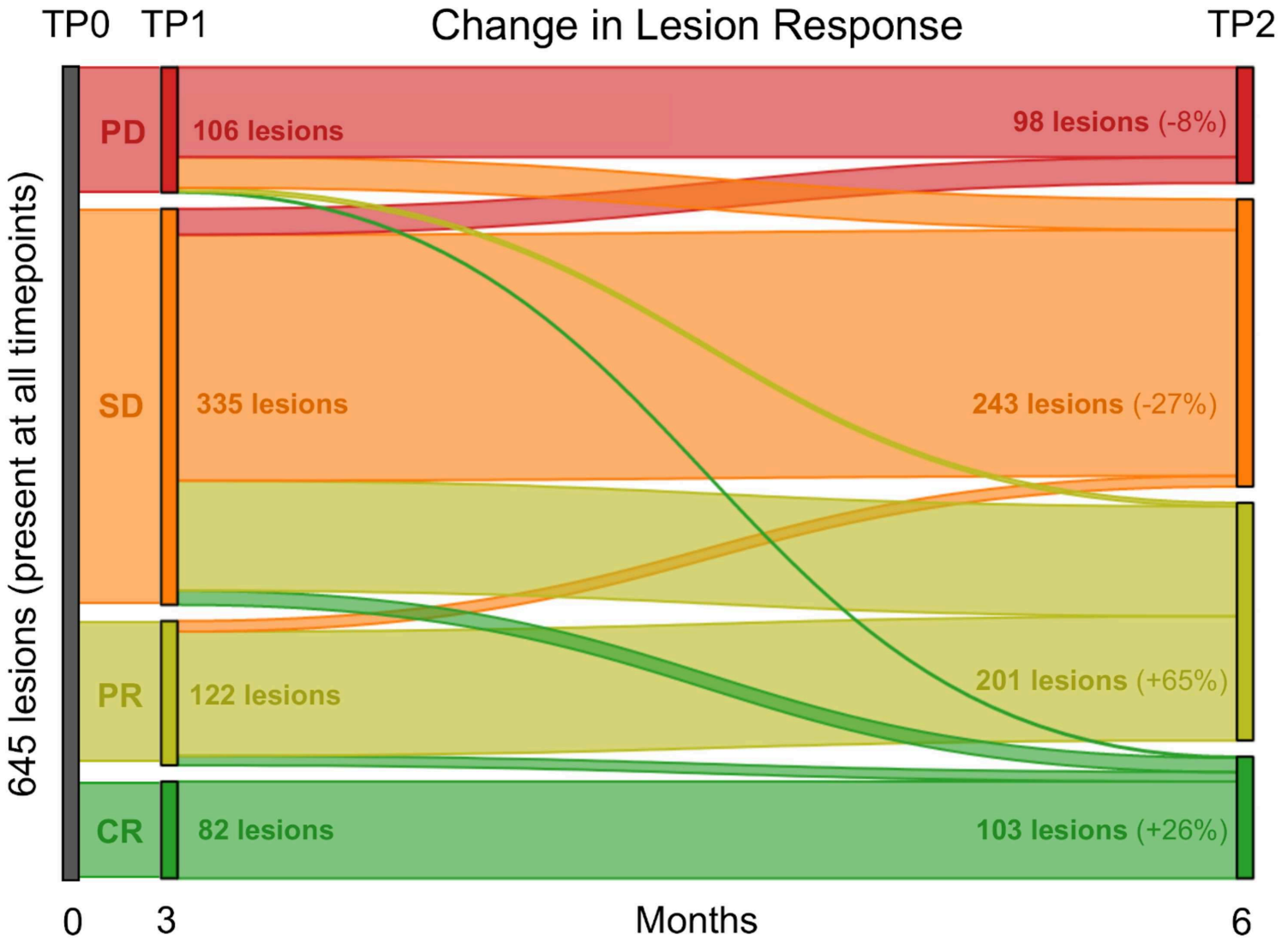

Fig. 2 Change of individual lesion response between all time-points Individual change of response between baseline (TP0), 3 months (TP1) and 6 months (TP2) for all lesions that were available at all 3 time-points (n = 645). 106 (16%) progressive lesions (PD) were identified at TP1, of which 30 changed to either complete response (CR), partial response (PR) or stable disease (SD) at TP2, representing pseudoprogression (PP). 76 (71.7%) lesions remained progressive throughout TP1 and TP2 and were classified as true progressive disease (TPD).

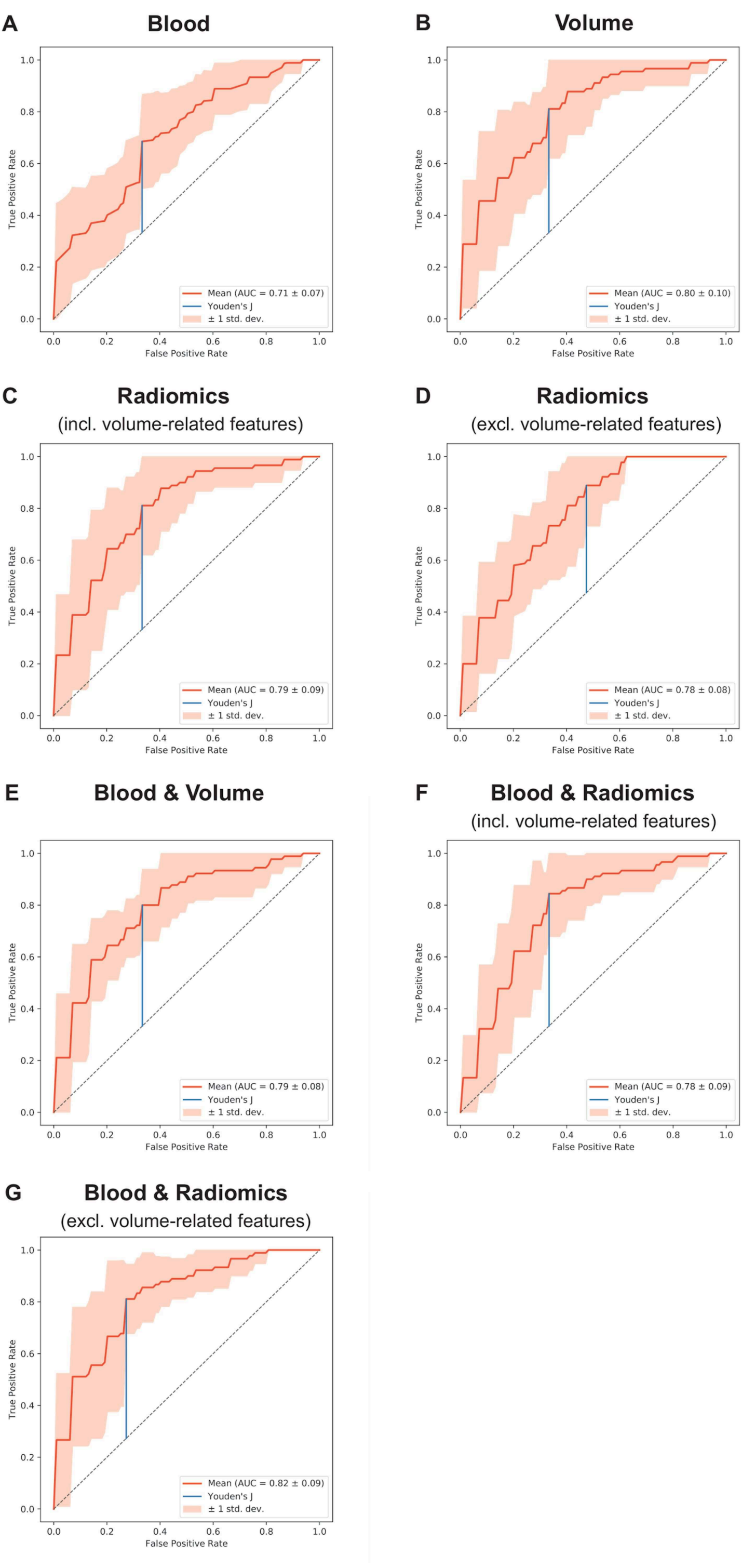

Fig. 3 Multivariate models for prediction of pseudoprogression. AUC curves for the best performing models of all seven model-classes. **A**. blood-based model. **B**. volume-based model. **C**. radiomics-based model (incl. volume-related features). **D**. radiomics-based model (excl. volume-related features). **E**. combined blood & volume-based model. **F**. combined blood & radiomics-based model (incl. volume-related features). **G**. combined blood & radiomics-based model (excl. volume-related features), which was the best performing model and achieved an AUC of 0.82 (sensitivity = 0.81 ± 0.13, specificity = 0.73 ± 0.35). This prediction model is based on the LDH level at TP1 and the relative change of CT coarseness between TP1 and TP0. Larger values of LDH and a larger decrease in CT coarseness indicated a lower chance of pseudoprogression.

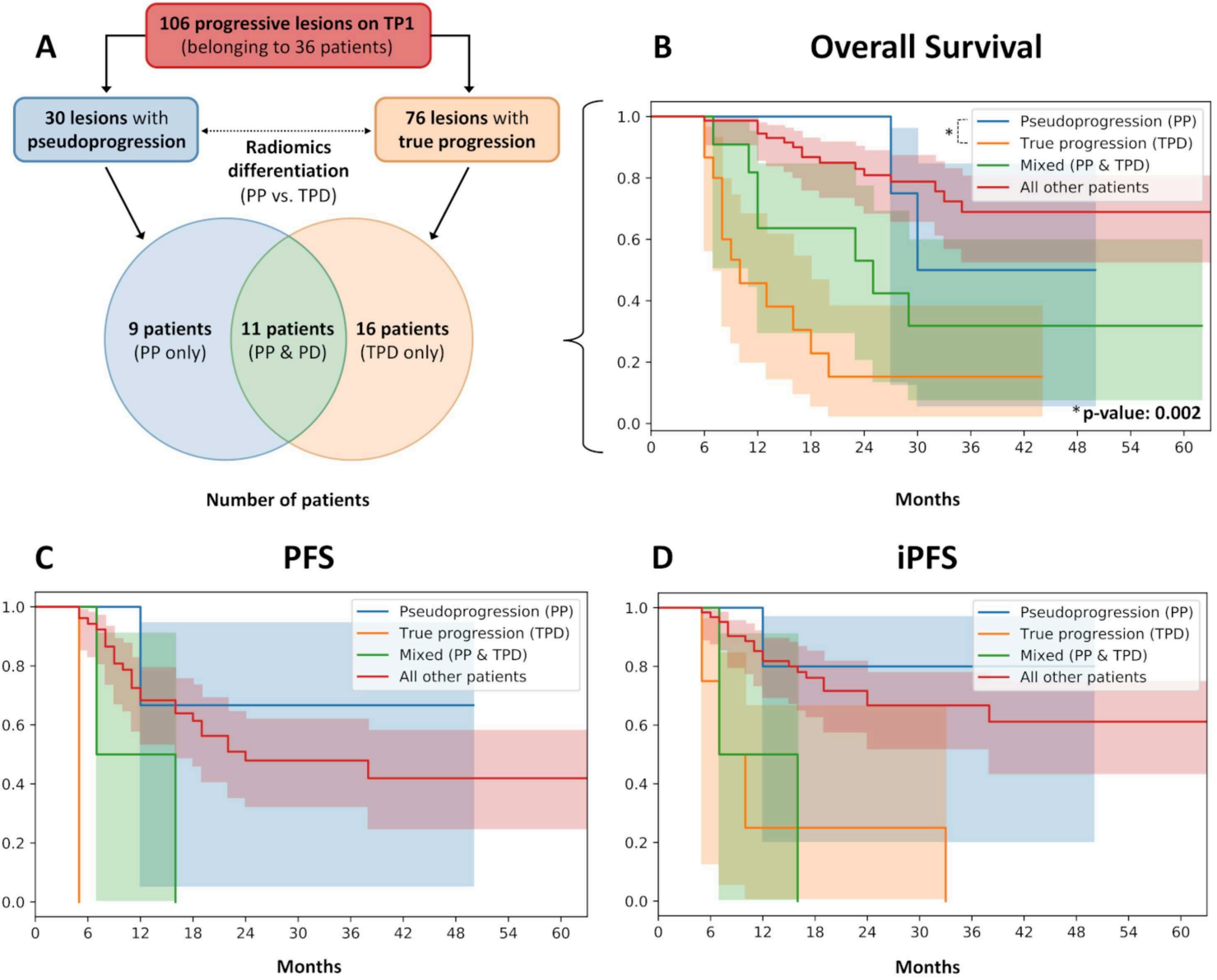

Fig. 4 Distribution of pseudoprogression and true progression and association with OS, PFS and iPFS A. 106 progressive lesions were present at TP1, 30 lesions with pseudoprogression were distributed across 20 patients with ≥1 PP-lesion. The 76 lesions with true progression were distributed across a total of 27 patients. An overlap of 11 patients presented with mixed PP&TPD lesions. Overall survival (B), progression-free survival (C) and immune-progression-free survival (D) of the different groups. The landmark was set at 5 months. PP-only patients had a significantly longer median OS of 30 vs. 10 months (p=0.002, FWER=0.010) compared to TPD-only patients with a 2-year OS of 100% vs. 15%.